Developing a Metadata Data Model for the National Health and Nutrition Examination Survey (NHANES)

Lewis E. Berman, MS, Computer Science, Centers for Disease Control and Prevention, Hyattsville, MD.
Allan L. Fisher, BS, Information Science, Orkand Corporation, Centers for Disease Control and Prevention, Hyattsville, MD.
Leighton Evans, BS, Information Processing, Orkand Corporation, Centers for Disease Control and Prevention, Hyattsville, MD.
Timothy Tilert, BS, Mathematics, Orkand Corporation, Centers for Disease Control and Prevention, Hyattsville, MD.

Abstract

The National Health and Nutrition Examination Survey (NHANES), developed by the Centers for Disease Control and Prevention (CDC), is a large and comprehensive health survey utilizing leading edge technologies to produce national estimates of health measures and the nutritional status of the U.S. population. Early NHANES metadata models grouped data by categories with little specificity and often not capturing the complexity of the survey. Subsequently, existing models at the Census Bureau, CDC, and the EPA were evaluated in addition to industry standards, such as DDI, Dublin Core, and ISO 1179. For the NHANES metadata model, the DDI standard and CDC Public Health Conceptual Model were chosen as the backbone for constructing the data model. The new model has led to increased data accuracy and several value-added products for producing codebooks, automatically checking questionnaire skip patterns, and producing questionnaire instrumentation.

Background

The fourth National Health and Nutrition Examination Survey (NHANES) took a new direction beginning in 1999. NHANES is designed to collect data that can be obtained by direct physical examination, clinical and laboratory tests, and related measurement procedures. This information is used to estimate either the prevalence of some disease or estimate the normative distribution of the characteristic in the total population. In previous NHANES surveys the metadata related to a survey data item was encapsulated in hardcopy formats and to a limited degree static electronic files. This logistics of dealing with these manuals or files prevents an analyst or researcher from easily determining the contextual information related to classes data items. To combat this problem CDC/NCHS explored a web-based searchable catalog of NHANES metadata.

An early version of this effort produced a prototype application using NHANES data grouped by data categories (e.g., abdomen, accident, allergy, etc), with each data category having multiple data items. The prototype successfully provided a proof-of-concept but exploited several weaknesses and limitations. Further research expanded this effort by incorporating metadata standards such as the Data Definition Initiative (DDI) standard, integration of the extensible markup language (XML) and XML tools, and the CDC Public Health Conceptual Model into a searchable on-line catalog of NHANES metadata.